Necessary libraries

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Loading dataset

```
dataset <- read_csv("preprocessed_data.csv")
```

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Pitanje 3: Uvjetuje li broj spavaćih soba cijenu kvadrata nekretnine?

Naš dataset ne sadržava cijenu po kvadratu, tj. kvadratnoj stopi (ft^2) za nekretninu, tako da ćemo tu vrijednost izračunati dijeljenjem cijene po kojoj je prodana sa stupcem GrLivArea, što je kvadratura područja za stanare.

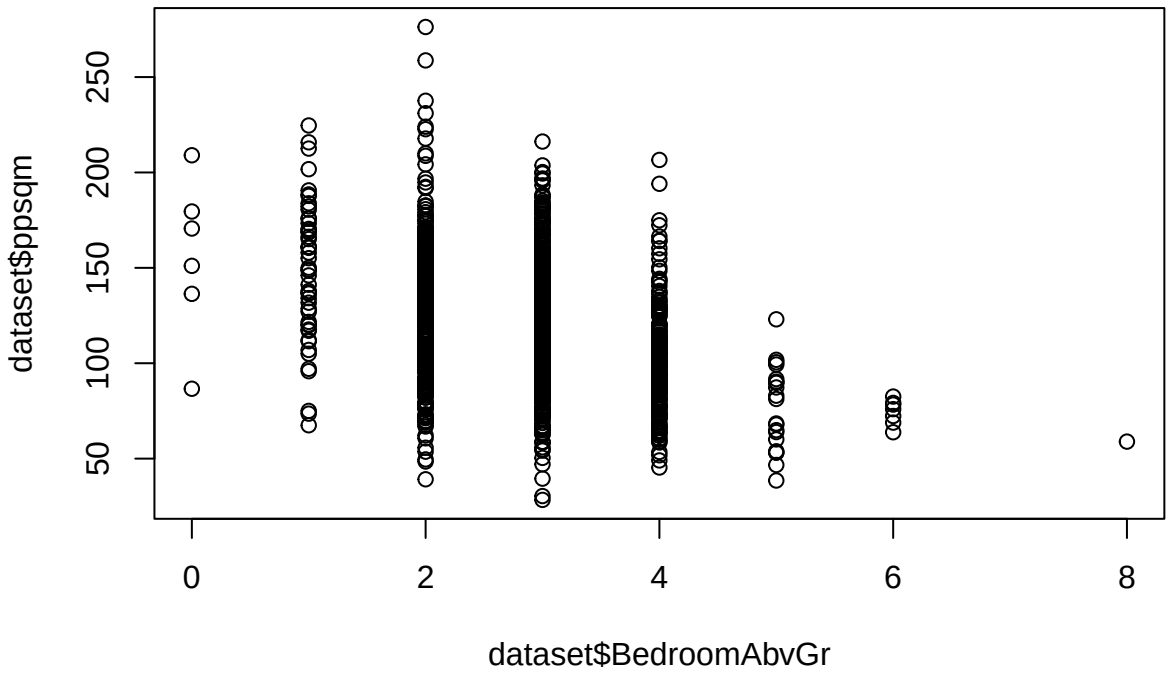
Za testiranje postavljamo hipoteze H_0 i alternativu H_1. Razmatramo utjecaj broja spavaćih soba u nekretnini na cijenu kvadrata nekretnine.

Analizom varijance testiramo:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{barem dvije sredine cijena po kvadratu nisu iste.}$$

Kako bismo dobili dojam o cijelom datasetu potrebnih podataka, plotat ćemo cijenu kvadrata ovisno o broju spavaćih soba u nekretnini.

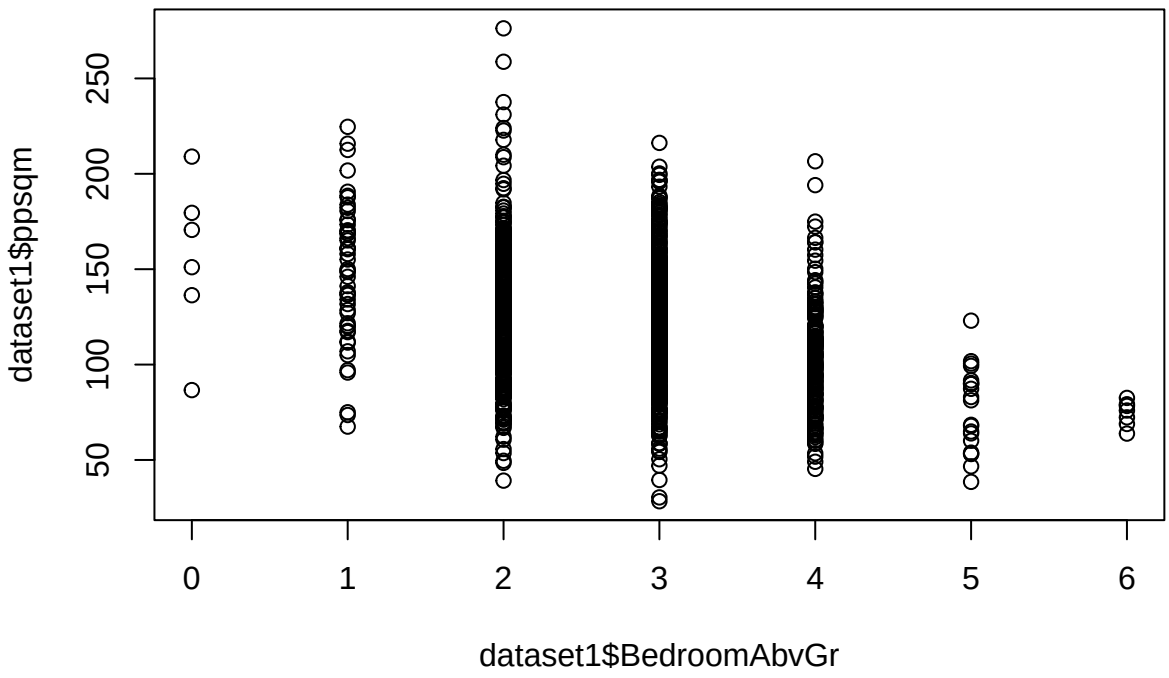
```
dataset$ppsqm = dataset$SalePrice / dataset$GrLivArea
plot(dataset$BedroomAbvGr, dataset$ppsqm)
```



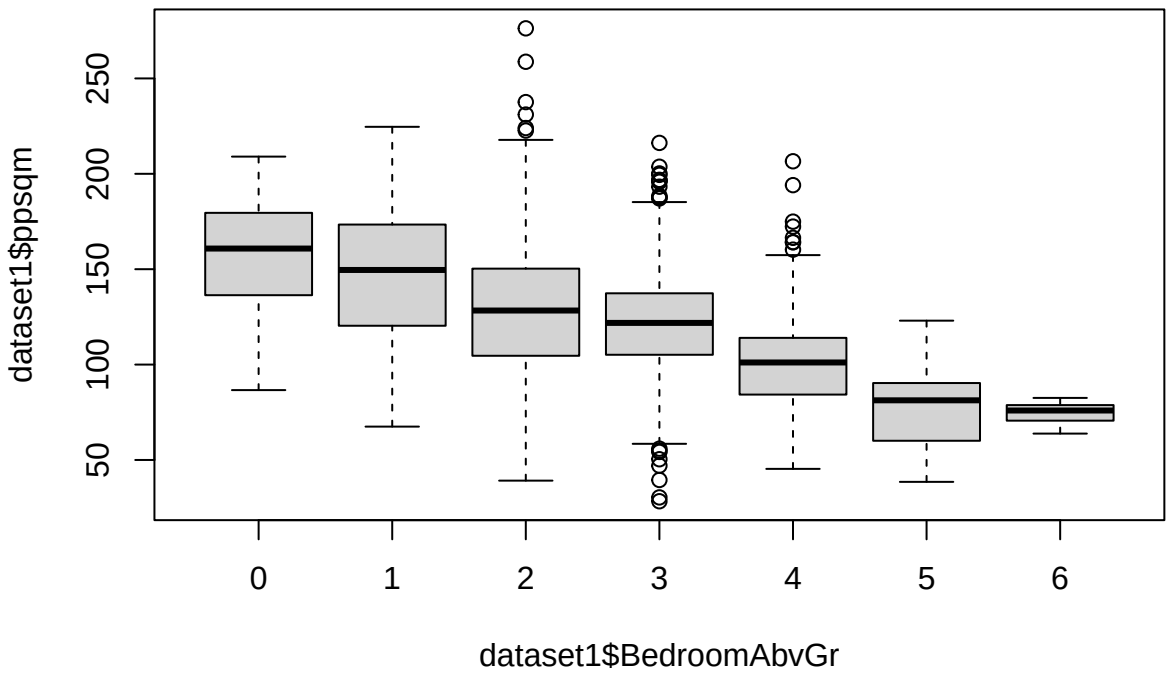
Ovdje možemo vidjeti distribuciju obzirom na cijenu kvadrata po broju spavaćih soba. Za provođenje testiranja mičemo stan sa 8 soba obzirom da imamo jednu vrijednost, što nam statistički ne pridonosi previše obzirom na malu veličinu uzorka.

Nakon toga napravit ćemo boxplot kako bismo dobili dojam o sredimama podkategorija po broju spavačih soba, te Q-Q Plot kako bismo procijenili normalnost cjelokupnog dataseta.

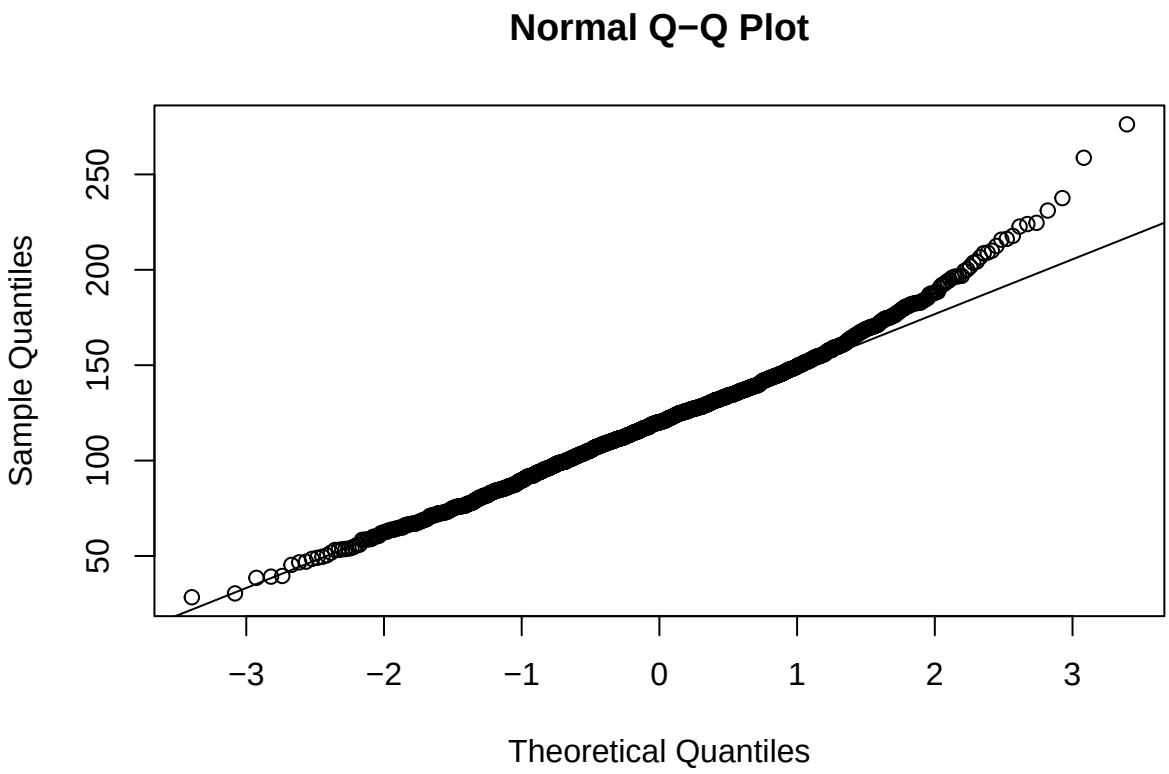
```
dataset1 = subset(dataset, BedroomAbvGr != 8)
plot(dataset1$BedroomAbvGr,dataset1$ppsqm)
```



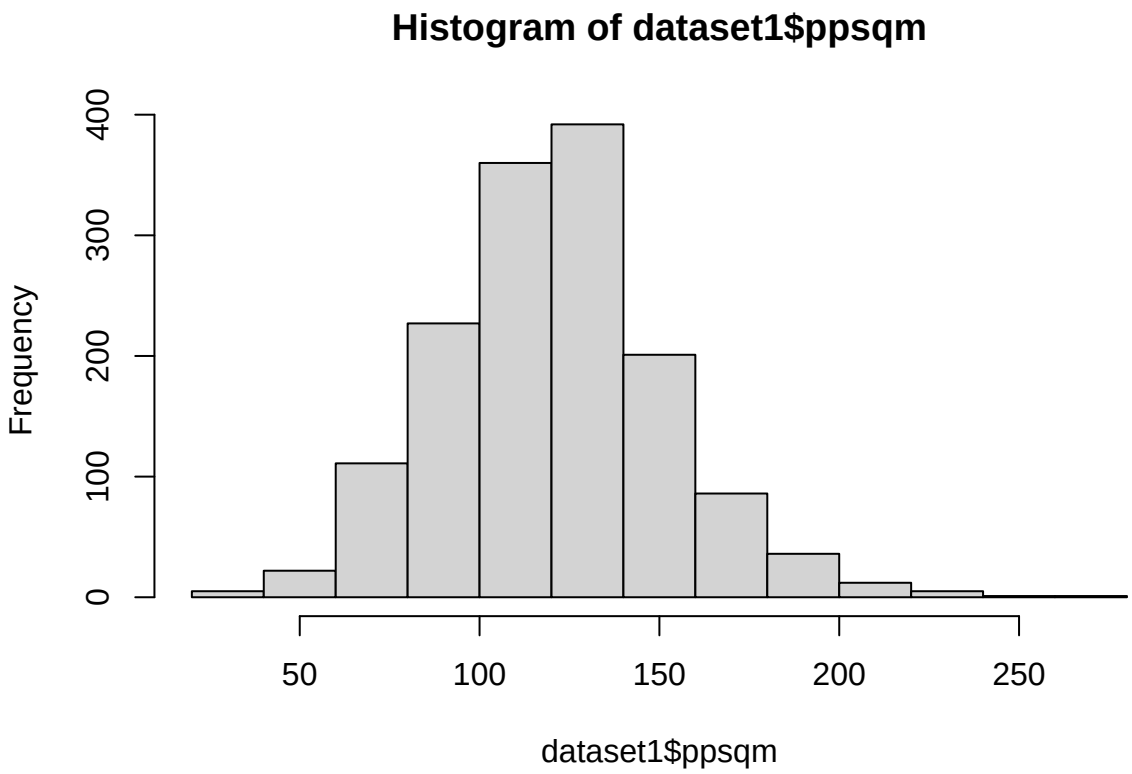

```
boxplot(dataset1$ppsqm ~ dataset1$BedroomAbvGr)
```




```
qqnorm(dataset1$ppsqm)
qqline(dataset1$ppsqm)
```

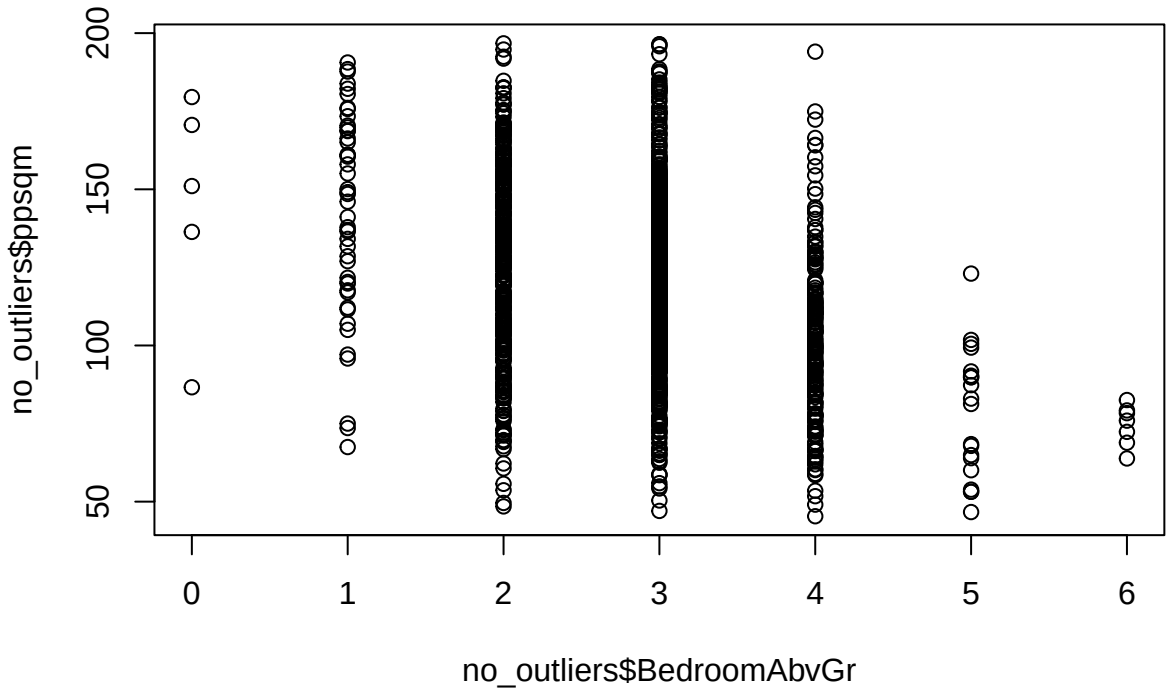



```
hist(dataset1$ppsqm)
```

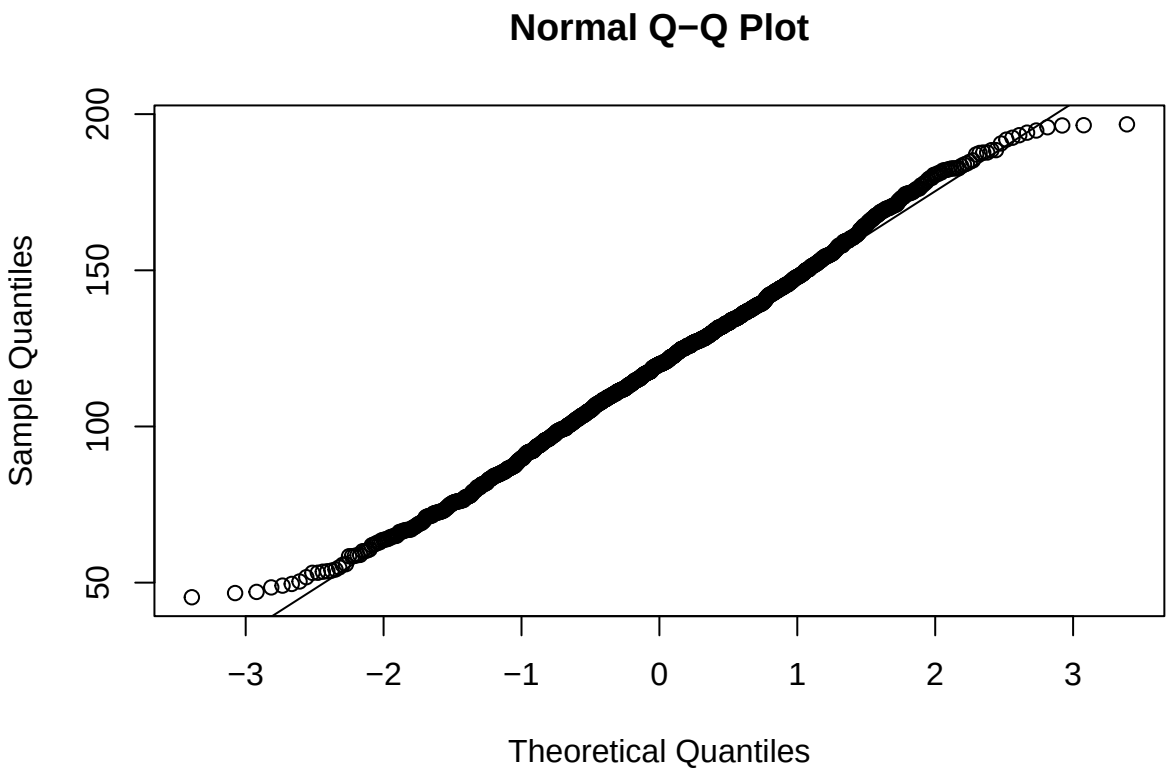


Vidimo da krajevi na Q-Q Plotu odstupaju od očekivane linije, stoga ćemo podatke očistiti od outliera tako da maknemo sve vrijednosti koje su veće od $Q3 + 1.5IQR$, te sve koje su niže od $Q1 - 1.5IQR$. Na histogramu također vidimo kako imamo distribuciju zakrivljenu udesno.

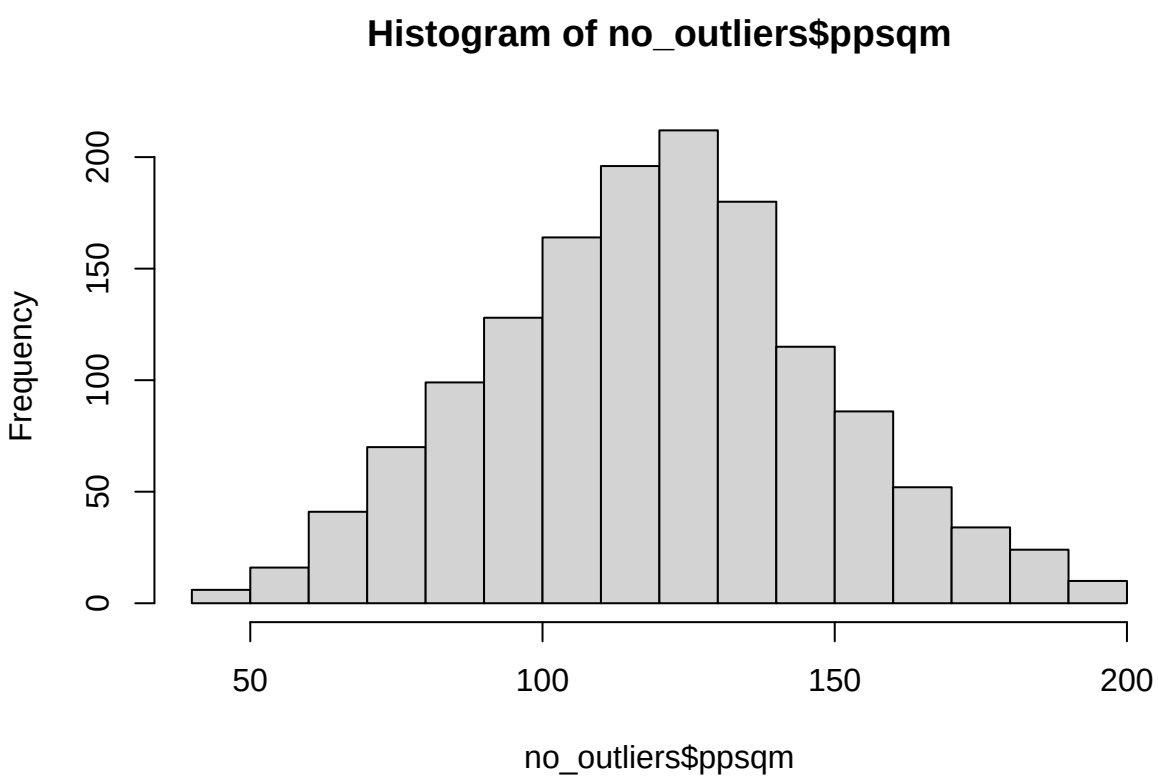
```
quartiles = quantile(dataset1$ppsqm, probs = c(.25, .75), na.rm=FALSE)
IQRppsqm = IQR(dataset1$ppsqm)

lower <- quartiles[1] - 1.5*IQRppsqm
upper <- quartiles[2] + 1.5*IQRppsqm

no_outliers = subset(dataset1, dataset1$ppsqm > lower & dataset1$ppsqm < upper)

plot(no_outliers$BedroomAbvGr, no_outliers$ppsqm)
```




```
qqnorm(no_outliers$ppsqm)
qqline(no_outliers$ppsqm)
```




```
hist(no_outliers$ppsqm)
```



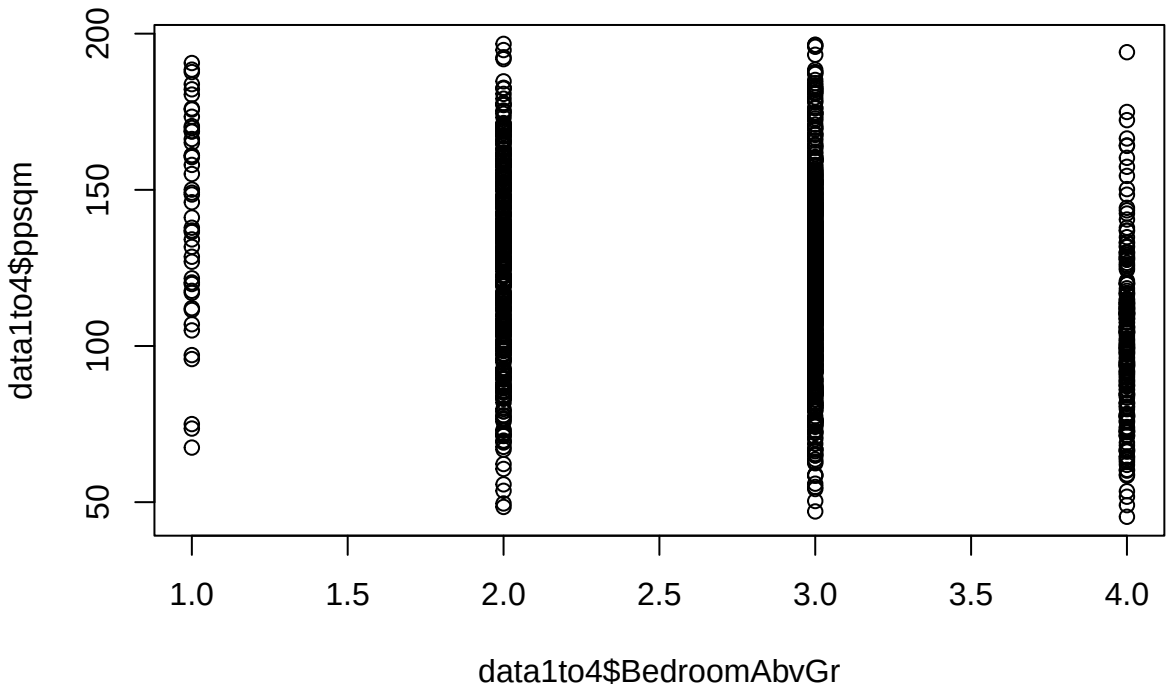
Prema Q-Q plotu dobili smo podatke koji su bolji od podataka prije čišćenja te zadovoljavaju pretpostavku normalnosti cijelog dataseta cijena po kvadratu. Sada i histogramom vizualiziramo distribuciju koja gotovo da ne izgleda zakrivljeno.

```
dataset2 <- dataset1[names(dataset1) %in% c('ppsqm', 'BedroomAbvGr')]

no_outliers %>%
  group_by(BedroomAbvGr) %>%
  count() -> dataset3

dataset3
```

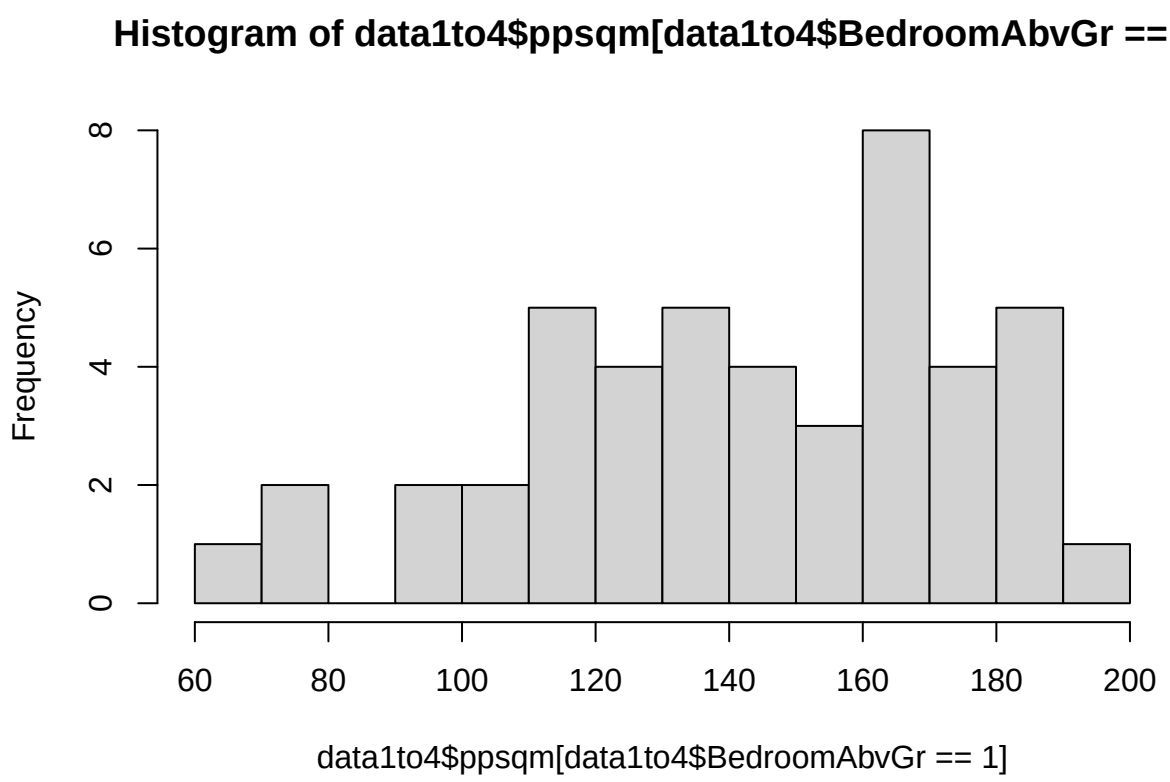
```
## # A tibble: 7 x 2
## # Groups:   BedroomAbvGr [7]
##   BedroomAbvGr     n
##          <dbl> <int>
## 1            0     5
## 2            1    46
## 3            2   347
## 4            3   796
## 5            4   212
## 6            5    20
## 7            6     7
```

Obzirom da stanove 0, 5 i 6 soba imamo malo podataka (između 10 i 100x manje od potkategorija s najvećim brojem podataka), njih nećemo uzeti u obzir za statističko testiranje, čime završavamo s podacima koji izgledaju ovako:

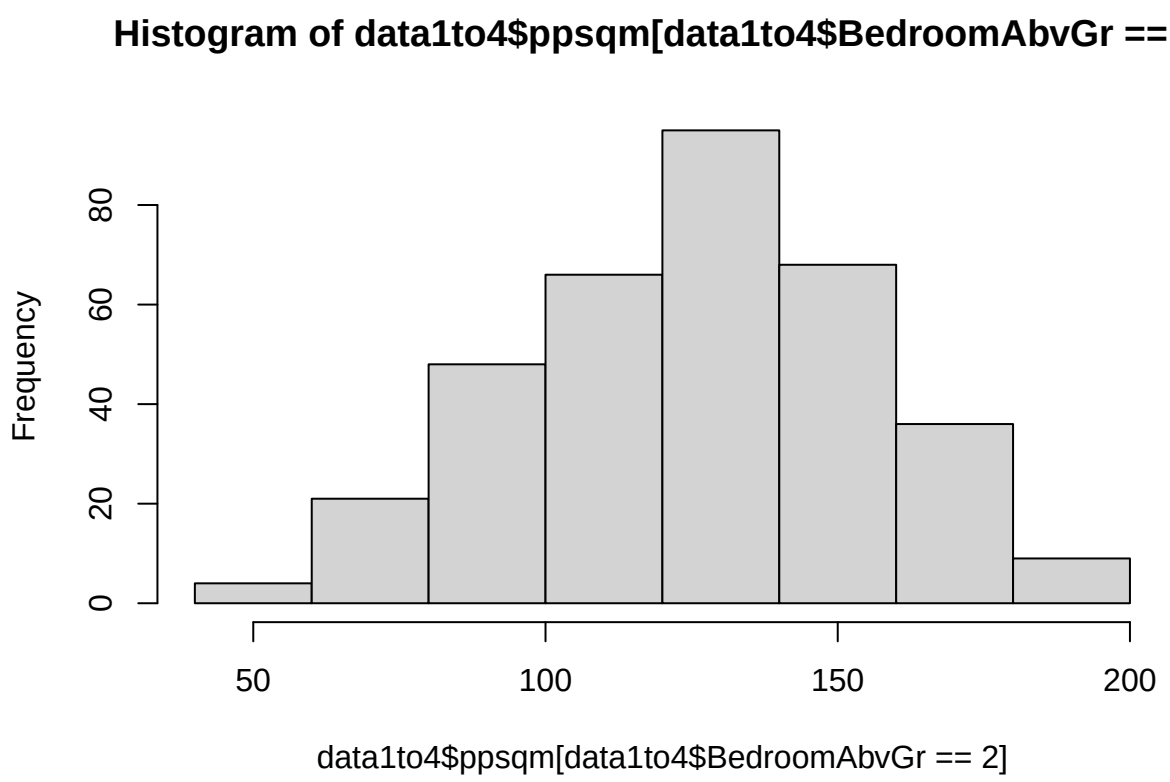
```r
data1to4 = subset(no_outliers, no_outliers$BedroomAbvGr > 0 & no_outliers$BedroomAbvGr < 5)

plot(data1to4$ppsqm ~ data1to4$BedroomAbvGr)
```



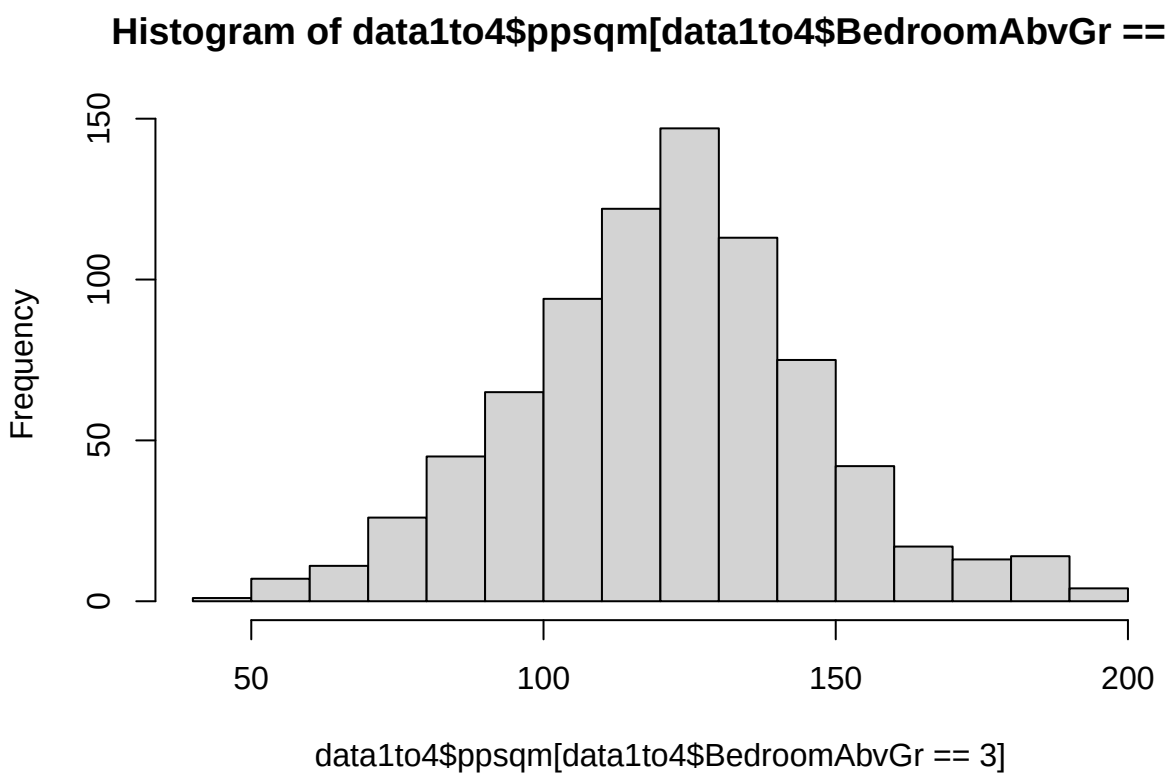

```r
hist(data1to4$ppsqm[data1to4$BedroomAbvGr == 1], breaks = 12)
```



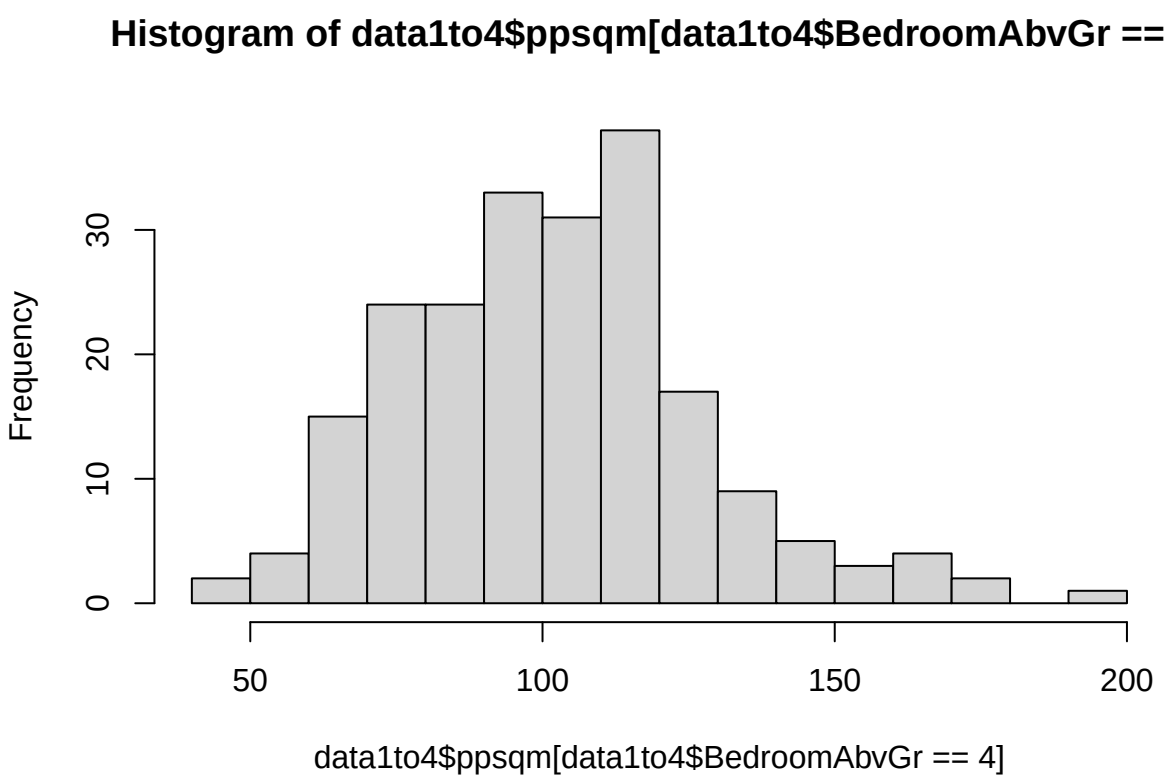

```
hist(data1to4$ppsqm[data1to4$BedroomAbvGr == 2])
```




```
hist(data1to4$ppsqm[data1to4$BedroomAbvGr == 3])
```




```
hist(data1to4$ppsqm[data1to4$BedroomAbvGr == 4], breaks = 20)
```




```
require(nortest)
```

```
## Loading required package: nortest
```

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(data1to4$ppsqm ~ data1to4$BedroomAbvGr)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data1to4$ppsqm by data1to4$BedroomAbvGr
## Bartlett's K-squared = 17.446, df = 3, p-value = 0.0005722
```

```
# Varijance
var_1room = var(data1to4$ppsqm[data1to4$BedroomAbvGr == 1])
var_2room = var(data1to4$ppsqm[data1to4$BedroomAbvGr == 2])
var_3room = var(data1to4$ppsqm[data1to4$BedroomAbvGr == 3])
var_4room = var(data1to4$ppsqm[data1to4$BedroomAbvGr == 4])

cat("Varijanca cijene po kvadratu stanova s 1 spavaćom sobom: ", var_1room, "\n")
```

```
## Varijanca cijene po kvadratu stanova s 1 spavaćom sobom:  1045.834
```

```r
cat("Varijanca cijene po kvadratu stanova s 2 spavaćom sobom: ", var_2room, "\n")
```

```
## Varijanca cijene po kvadratu stanova s 2 spavaćom sobom:  900.2783
```

```r
cat("Varijanca cijene po kvadratu stanova s 3 spavaćom sobom: ", var_3room, "\n")
```

```
## Varijanca cijene po kvadratu stanova s 3 spavaćom sobom:  652.8112
```

```r
cat("Varijanca cijene po kvadratu stanova s 4 spavaćom sobom: ", var_4room, "\n")
```

```
## Varijanca cijene po kvadratu stanova s 4 spavaćom sobom:  652.5645
```

Iako nam Bartlettov test sugerira da varijance između poduzoraka soba sa 1 do 4 spavaćih soba nisu homogene, vidimo da su istog reda veličine, stoga nastavljamo sa testiranjem podataka.

```r
lillie.test(data1to4$ppsqm)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data1to4$ppsqm
## D = 0.017688, p-value = 0.3579
```

```r
lillie.test(data1to4$ppsqm[data1to4$BedroomAbvGr == 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data1to4$ppsqm[data1to4$BedroomAbvGr == 1]
## D = 0.10279, p-value = 0.2586
```

```r
lillie.test(data1to4$ppsqm[data1to4$BedroomAbvGr == 2])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data1to4$ppsqm[data1to4$BedroomAbvGr == 2]
## D = 0.041909, p-value = 0.1456
```

```r
lillie.test(data1to4$ppsqm[data1to4$BedroomAbvGr == 3])
```

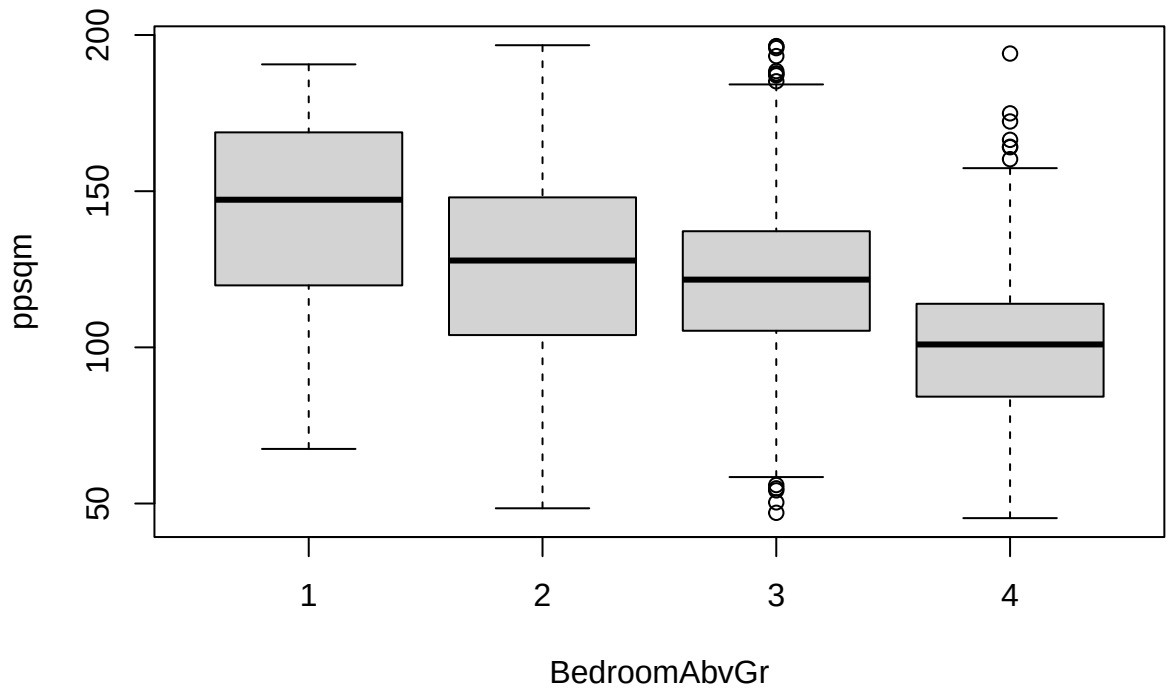
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data1to4$ppsqm[data1to4$BedroomAbvGr == 3]
## D = 0.02798, p-value = 0.1365
```

```
lillie.test(data1to4$ppsqm[data1to4$BedroomAbvGr == 4])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data1to4$ppsqm[data1to4$BedroomAbvGr == 4]
## D = 0.069636, p-value = 0.01426
```

P-value dobiven Lillieforsovim testom za svaku potkategoriju nam sugerira normalnost. Ovaj test smo koristili kao "sanity check", kako bismo vidjeli da nismo donijeli neke zaključke potpuno krivo. ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavki normalnosti i homogenosti varijance, no ipak smo proveli testiranja kako bi smo vidjeli kolika su stvarno ta odstupanja.

```
boxplot(ppsqm ~ BedroomAbvGr,data=data1to4)
```



Grafički nam prikaz sugerira da postoji razlika u sredini cijene po kvadratu između potkategorija nekretnina.

Sada ćemo provesti test ANOVA-e nad setom podataka.

```
aov = aov(data1to4$ppsqm ~ data1to4$BedroomAbvGr)

summary(aov)
```

```
##                          Df  Sum Sq Mean Sq F value Pr(>F)
## data1to4$BedroomAbvGr     1   89864   89864   121.7 <2e-16 ***
```

```
## Residuals             1399 1033254     739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obzirom na dobiveni p-value, na razini značajnosti od $\alpha = 0.05$ odbacujemo hipotezu H_0 u korist alternativne hipoteze H_1, dakle odbacujemo hipotezu da su sredine uzoraka jednake.

Sada želimo procijeniti model koji bi nam pomoću varijable o broju spavaćih soba u nekretnini objasnio cijenu kvadrata iste te nekretnine.

sad treba fittat lineranu reg za model